Attorney's Docket No. K&A 21-0851

APPLICATION

FOR UNITED STATES LETTERS PATENT

SPECIFICATION

TO ALL WHOM IT MAY CONCERN:

BE IT KNOWN THAT WE, **WILLIAM J. BUSHEE**, a citizen of UNITED STATES OF AMERICA, and **THOMAS W. TIAHRT**, a citizen of UNITED STATES OF AMERICA, and **MICHAEL K. BERGMAN**, a citizen of the UNITED STATES OF AMERICA, have invented a new and useful **METHOD AND SYSTEM FOR AUTOMATIC HARVESTING AND QUALIFICATION OF DYNAMIC DATABASE CONTENT** of which the following is a specification:

METHOD AND SYSTEM FOR AUTOMATIC HARVESTING AND QUALIFICATION OF DYNAMIC DATABASE CONTENT

BACKGROUND OF THE INVENTION

Incorporation By Reference

This patent application discloses an invention of a system integrating multiple constituent systems. These constituent systems are disclosed and described in the following co-pending patent applications, all of which are subject to an obligation of assignment to the same person. The disclosures of these applications are herein incorporated by reference in their entireties.

METHOD FOR AUTOMATIC SELECTION OF DATABASES FOR SEARCHING, William J. Bushee, Filed July ___, 2001, Application Serial Number ___________.

AUTOMATIC SYSTEM FOR CONFIGURING TO DYNAMIC DATABASE SEARCH FORMS, William J. Bushee, Filed July ___, 2001, Application Serial Number ___________.

SYSTEM AND METHOD FOR EFFICIENT CONTROL AND CAPTURE OF DYNAMIC DATABASE CONTENT, William J. Bushee and Thomas W. Tiahrt, Filed July ___, 2001, Application Serial Number ___________.

SYSTEM FOR AUTOMATICALLY CATEGORIZING CONTENT IN HIERARCHICAL SUBJECT STRUCTURES, Thomas W. Tiahrt, Michael K. Bergman, and William J. Bushee, Filed July ___, 2001, Application Serial Number

SYSTEM AND METHOD FOR FLEXIBLE INDEXING OF DOCUMENT CONTENT, Thomas W. Tiahrt, Filed July ___, 2001, Application Serial Number ___________.

SYSTEM FOR AUTOMATICALLY CREATING SYNTHETIC SUMMARIES FROM DOCUMENT CONTENT, Thomas W. Tiahrt, William J. Bushee, and Michael K. Bergman, Filed July ___, 2001, Application Serial Number ___________.

Field of the Invention

The present invention relates to search engines and database searching techniques and more particularly pertains to a new method and system for automatic harvesting and qualification of dynamic database content for efficiently providing highly relevant and timely information in response to user's queries.

Description of the Prior Art

Many enterprises, whether business, governmental, or any other organized undertaking, require large amounts of "current" information to be analyzed and available for use in the daily execution of their activities. Often the informational needs of the enterprise can be categorized into discrete subject areas or domains. Each of these domains may have additional divisions providing increasing granularity or specificity of the subject matter.

Since its inception, the Internet has held the promise of real-time access to an almost inexhaustible supply of information, stored on computers throughout the world, in near real time. However, sorting through the information available to find documents relevant to a given question or query can be laborious; and a method to speed this process was needed. Search engines are known in the prior art and allow a user to search for sites that have some keyword corresponding to the user's query. While it is true that millions of documents are readily available as static pages to users through search engines, much more of the total content of the Internet, in the form of dynamic content, has remained relatively difficult to access through more conventional search engine techniques. For the purpose of clarity, a static page of a network database provides the same content to virtually every user accessing the database, usually in the form of the same document or page (or set of documents or pages). A dynamic network database presents dynamic content to each user accessing the database, and the dynamic content usually comprises unique documents or pages that are in response to and are based at least in part on the user's query.

The dynamic content, while available, often requires independent knowledge of the exact location of the document, sophisticated search techniques, or in many cases the use of professional researchers to attempt to "mine" the needed information.

Additionally, the resources required to evaluate all of the information identified by a conventional search engine in order to filter out non-relevant information can be more than substantial. The resources used may include, by way of example and not limitation, transmission bandwidth, data storage, and time (both of system usage and of personnel) required to filter out related but not relevant information. The need to capture and organize relevant information can be overwhelming, and an automated system is required to effectively solve this problem.

In these respects, the method and system for automatic harvesting and qualification of dynamic database content according to the present invention substantially departs from the conventional concepts and designs of the prior art, and in so doing provides a system primarily developed for the purpose of efficiently providing highly relevant and timely information in response to user's queries.

SUMMARY OF THE INVENTION

In view of the foregoing disadvantages inherent in the known types of search engines and database searching techniques now present in the prior art, the present invention provides a new method and system for automatic harvesting and qualification of dynamic database content construction wherein the same can be utilized for efficiently providing highly relevant and timely information in response to user's queries.

The invention contemplates a method for the automatic harvesting and qualification of dynamic database content. The method may include obtaining an initial categorization structure for organizing a plurality of subject areas of information, obtaining a plurality of parametric information lists for optimizing operation to a user's requirements, acquiring a listing of a plurality of qualified databases from said candidate database listing by matching each one of a candidate databases to said plurality of subject areas, obtaining a query from the user, said query being associated with a subject area, submitting said query to said plurality of qualified databases, acquiring a collection of responsive content from said plurality of qualified databases, indexing said responsive content to form an index of facilitating searching said collection of responsive content, and publishing a summary of said collection of responsive content for review by the user. The invention also contemplates a system for carrying out the various aspects of the method.

There has thus been outlined, rather broadly, the more important features of the invention in order that the detailed description thereof that follows may be better understood, and in order that the present contribution to the art may be better appreciated. There are additional features of the invention that will be described hereinafter and which will form the subject matter of the claims appended hereto.

In this respect, before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not limited in its application to the details of construction and to the arrangements of the components set forth in the following description or illustrated in the drawings. The invention is capable of other embodiments and of being practiced and carried out in various ways. Also, it is to be understood that the phraseology and terminology employed herein are for the purpose of description and should not be regarded as limiting.

As such, those skilled in the art will appreciate that the conception, upon which this disclosure is based, may readily be utilized as a basis for the designing of other structures, methods and systems for carrying out the several purposes of the present invention. It is important, therefore, that the claims be regarded as including such equivalent constructions insofar as they do not depart from the spirit and scope of the present invention.

The objects of the invention, along with the various features of novelty which characterize the invention, are pointed out with particularity in the claims annexed to and forming a part of this disclosure. For a better understanding of the invention, its operating advantages and the specific objects attained by its uses, reference should be made to the accompanying drawings and descriptive matter in which there are illustrated preferred embodiments of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be better understood and objects other than those set forth above will become apparent when consideration is given to the following detailed description thereof. Such description makes reference to the annexed drawings wherein:

Figure 1 is a schematic functional interconnect view of a new method and system for automatic harvesting and qualification of dynamic database content according to the present invention.

Figure 2 is a schematic functional flow diagram view of the present invention.

Figure 3 is a schematic functional flow diagram of the selection module view of the present invention.

Figure 4 is a schematic functional flow diagram of the results index view of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

The system for automatic harvesting and qualification of dynamic database content of the invention (see Figures 1 through 4) performs a plurality of major functions, which may include acquisition of databases to be queried, acquisition of dynamic content in response to the query, indexing the dynamic content, and publication of the results. The system's major modes of operation include an initial capture of dynamic content which is referred to as a “harvest”. Additionally a query servicing mode is also incorporated into the system. The harvest will be described in the following several paragraphs. The query servicing mode will be described in terms of differences from the initial harvest. It is noted that the system works equally well with static content databases, but the full advantages of the system are exploited when working with dynamic content databases. From this point forward the term “content” is assumed to encompass both dynamic content as well as static content.

The system obtains an initial listing of databases, an initial categorization structure defining the information domain, and a plurality of parametric information lists. The system begins the acquisition of databases by matching the query or queries to the database to provide content which is highly relevant to the query or queries. The term query as used herein is presumed to include one or more queries.

The system uses a first one of the parametric information lists is a candidate database list, which provides an extensive group of candidate databases to be considered. The candidate databases can extend into the tens of thousands to hundreds of thousands. For example, on the Internet today, it is estimated there may be perhaps on the order of 250,000 searchable dynamic databases.

An initial page presented by each candidate database is evaluated for relevance to the specific domain and sub-classification of information or subject area. Any database which is determined to not be relevant to the subject area is removed from consideration for that subject area. A number of the remaining databases are selected for further consideration. The specific number of databases selected may be limited by a user-defined parameter (such as a database relevancy parameter), which establishes a minimum threshold of relevancy for any given subject area.

Each of the selected databases may have a unique set of requirements for submitting queries and retrieving documents. In order to facilitate the efficient harvest of content, each of the selected databases is analyzed for these requirements and a configuration file is created. For each database, the configuration file may serve as a translator between a generic query established by the user and the unique requirements of each database. The configuration file provides the system with information for the proper submission of queries and retrieval of responses for each one of the selected databases.

Each of the selected and configured databases is then again evaluated for relevance to the subject area. A sample query from the subject area is submitted to each of the selected databases. Responsive pages or documents are then gathered from each of the databases. These responsive documents are evaluated for relevance to the subject area. Each of the databases is assigned a numerical score representing relevance to the subject area. Databases with a sufficiently high numerical score are then qualified for use in the subject area. A different collection of databases may be qualified for each subject area. The qualified databases are then used for the next major function: document acquisition.

The system uses the qualified databases and the initial categorization structure (such as a collection of subject matter areas) along with a series of queries to perform an initial harvest of content. The queries are queued and submitted to qualified databases. The responsive content from each database is captured and stored in a central location.

A difference between the initial harvest and the query servicing modes occurs at this point in the overall process. In an initial harvest the responsive content is captured or downloaded from the qualified database. In the query servicing mode, the central location is checked for the document before resorting to downloading the document from the source database. If the central location has a current copy of the document, the systems resources are not used to download a new copy from the source database.

The system next performs the major function of indexing the content for facilitating searching of the content. Here again is a difference between the initial harvest and query servicing modes. The index is created for documents qualified after the initial harvest. The index is used to find content matching a query during the query servicing mode.

The system parses each piece of content into constituent words for processing. The system then compares each of the words to a fourth one of the parametric list (such as a stop list). A stop list contains terms which have been determined not to add value to the index, and therefore these terms are not processed. Each word, which is not on the stop list, is then stemmed into its base prefix (such as a stem word) to facilitate efficient indexing. The location of each stem word in every piece of content is then recorded in the index, such that a user can search for any term based upon its corresponding stem word throughout the entire collection of content or documents through the index.

A summary of each piece of content may be created if a summary was not provided by the qualified database. The summary may provide a listing of keywords relevant to the subject area, or an extract of a particularly relevant portion of the piece of content. This is especially important for content taken from large databases of documents (such as, for example, patent databases) where summaries for each document are typically not provided or available.

As a final step in the indexing process, the system records a plurality of statistics associated with each piece of content. Illustratively, the plurality of statistics may include, but is not limited to: the title of the piece of content, the number of internal links in the piece of content, the number of external links in the piece of content, the number of terms in the piece of content, the length of the piece of content, the database which provided the piece of content, and whether the content was static or dynamic.

The indexing operation may also include recording a set of statistics describing the collection of content as a whole. In a preferred embodiment these statistics may include the number of pieces of content, the average number of terms per piece of content, the standard deviation of the number of terms, the total number of bytes to store the collection of content, and the total number of terms in the collection of content.

After all of the queries have been submitted to the qualified databases and the responsive content has been captured and stored in a central location, the system matches each piece of responsive content to the initial categorization structure. The initial categorization structure is a tree configuration with each domain being a first level of classification and each sub-classification being a branch depending from the first level of classification or another sub-classification. After this match has been performed, the system filters the categorization structure. This filtering may include a check for duplicate documents matched to the same classification, limiting the number of documents matched to any one classification or sub-classification based on a user defined parameter (such as a population parameter), and limiting the number of classifications or sub-classifications to which any one piece of content may be matched, based on a user defined parameter (such as an occurrence parameter). Additionally, the system may use a second parametric listing (such as an exclusion list) and a third parametric listing (such as an inclusion list) to inhibit matches or restrict matches (respectively) based upon a predetermined listing of terms and database sources for each subject area. After the filtering is complete, a categorization file is created which records the matches of the stored copies of the responsive content for each subject area.

Finally, the system generates pages facilitating the recall of any piece of content in associate with a user's query. The user may submit a query to the system. The system will then match the query to the harvested content and return a page providing a listing of each relevant piece of content in the collection, along with a summary of the piece of content.

Therefore, the foregoing is considered as illustrative only of the principles of the invention. Further, since numerous modifications and changes will readily occur to those skilled in the art, it is not desired to limit the invention to the exact construction and operation shown and described, and accordingly, all suitable modifications and equivalents may be resorted to, falling within the scope of the invention.